Discovering the relevant variables in a large clinical database by back-fitting fixed effects in a mixed linear model: Study of a long-term electrophysiological survey of cochlear implanted patients

Rafael Laboissière, Michel Mazzuca, Hung Thai-Van, Lionel Collet

Discovering the relevant variables in a large clinical database by back-fitting fixed effects in a mixed linear model: Study of a long-term electrophysiological survey of cochlear implanted patients

Rafael Laboissière[a], **Michel Mazzuca**[a], **Hung Thaï-Van**[a], & **Lionel Collet**

[a]Groupe Parole, Audiologie, Communication et Santé
Centre de Recherche en Neurosciences de Lyon
⟨rafael.laboissiere@inserm.fr⟩

Mots clefs : auditory evoked potentials, maturation, large data set, mixed-models

Clinical surveys are routine procedures that are typically run in Hospital services and intended to evaluate patients after surgery or therapeutic treatment. They differ from clinical trials, where specific drugs or new therapies are investigates according to a planned protocol and with a predetermined cohort. They also differ from basic research experiments, which are done in even narrowly controlled setups in laboratories. Clinical surveys, on the other hand, can include a huge amount of patients, can last for a very extended amount of time (sometimes even without a fixed time to be terminated) and can include a large amount of independent factors and dependent measured variables.

In this paper, we present the analysis of a large data set collected on cochlear implanted patients at the Audiology Service of the Édouard Herriot Hospital in Lyon. The analysis on a reduced fraction of this data set has already been published elsewhere (Thai-Van et al., 2007). A cochlear implant (CI) is a surgically implanted electronic device that restores the sense of audition to profoundly deaf patients. It acts as a transducer between the sound captured by an external microphone, placed closely to the patient's ear, and the neural cells at the extremity of the auditory nerve. The transducer device is an linear array of electrodes that is introduced in the scala tympani of the cochlea. After surgery, the patients do systematic visits to the hospital service, during which electrophysiological tests are carried out in order to verify the transmission of the auditory information from the implant to the brain. In these tests, specific electrodes are activated and the evoked potentials of the brainstem auditory relays are measured using scalp electrodes (Guiraud et al., 2007).

The first obstacle when trying to analyze such a data set comes from the fact that the information are scattered through medical records and are not always in a format that a statistician would expect them to be. A preliminary work of data formatting and quality control must then be done. Already at this point of the study, the R software presents itself as a powerful, if not essential, tool. The data set can be organized according to several independent factors, namely: the sex of the patient (male of female), the age at implantation (in days), the duration of privation before implantation (in days), the side of the implanted ear (left or right), the number of the activated electrode (in our case, #5 and #20, respectively more basal or more apical). Also, the patients are followed longitudinally and the main independent variable in the study is the duration of CI use at the date of visit (in days). The dependent measures are the latency of the waves III (response from the superior olive) and V (response from the inferior colliculus). The main goal of the study is to evaluate the neuronal maturation of the auditory pathways with the duration of CI use and how this maturation depends on the considered independent factors.

The data set included 232 patients (112 females and 120 males), some of them implanted bilaterally. The visits dates range from one month after the surgery until 14 years and a total

of over 13,000 independent measures are present in the data set. The data is well suited for a linear mixed-model analysis (Bates et al., 2011), in which the intercept latency for each patient is considered as a random factor. When the analysis is performed with an omnibus model including all the factors and their interactions, we end up with 43 factors and the analysis become almost unfeasible. In order to retain the factors that really matter in explaining the data, we back-fitted the fixed effects using the LMERConvenienceFunctions package of R (Newman et al., 2011; Tremblay and Tucker, 2011). This technique estimates the F statistics of each factor for the fitted model and the algorithm proceeds from the highest-order interaction terms towards the individual factors. The factor with the smallest value of F is considered for removal. A statistical test is then made between the models with and without this factor by using the log-likelihood ratio test. If the resulting p-value is below a given threshold α, then the factor is kept, otherwise it is discarded. This is done progressively until there are no factors left. Finally, an ANOVA table of the reduced model is computed using conservative estimations of the degrees of freedom of the denominator.
Using this technique we could find evidence for the following effects: (1) a difference in the response of electrodes #5 and #20, related to the anatomical position in the cochlear modiolus; (2) a latency difference between males and females, that correlates with anthropometric data; (3) a fast maturational rate followed by a standing plateau and a later increase of the latencies along the duration of CI use; (4) a lack of later increase in the latency time with CI duration use when the interval III–V is considered; and (5) a difference in the behavior of right *vs.* left implanted ears that interacts with the age at implantation. This later effect, the most interesting one found in thiss study, was confirmed in a subset of the data, where patients were selected to form matched groups in age and ear side.

References

Bates D, Maechler M, and Bolker B (2011) *lme4: Linear mixed-effects models using S4 classes.* R package version 0.999375-41 (URL: `http://CRAN.R-project.org/package=lme4`).

Guiraud J, Besle J, Arnold L, Boyle P, Giard MH, Bertrand O, Norena A, Truy E, and Collet L (2007) Evidence of a tonotopic organization of the auditory cortex in cochlear implant users. *J Neurosci* 27(29): 7838–7846.

Newman AJ, Tremblay A, Nichols ES, Neville HJ, and Ullman MT (2011) The influence of language proficiency on lexical semantic processing in native and late learners of English. *J Cogn Neurosci* PMID #21981676.

Thai-Van H, Cozma S, Boutitie F, Disant F, Truy E, and Collet L (2007) The pattern of auditory brainstem response wave V maturation in cochlear-implanted children. *Clin Neurophysiol* 118(3): 676–89.

Tremblay A and Tucker BV (2011) The effects of n-gram probabilistic measures on the processing and production of four-word sequences. *The Mental Lexicon* (accepted for publication).